A Multifaceted Study on Eye Contact based Speaker Identification in Three-party Conversations

Yu Ding
University of Houston
Houston, TX, USA
yding7@uh.edu

Yuting Zhang
University of Houston
Houston, TX, USA
yzhang115@uh.edu

Meihua Xiao
East China Jiaotong University
Nanchang, Jiangxi, China
xiaomh@ecjtu.edu.cn

Zhigang Deng
University of Houston
Houston, TX, USA
zdeng4@uh.edu

ABSTRACT
To precisely understand human gaze behaviors in three-party conversations, this work is dedicated to look into whether the speaker can be reliably identified from the interlocutors in a three-party conversation on the basis of the interactive behaviors of eye contact, where speech signals are not provided. Derived from a pre-recorded, multimodal, and three-party conversational behavior dataset, a statistical framework is proposed to determine *who is the speaker* from the interactive behaviors of eye contact. Additionally, with the aid of virtual human technologies, a user study is conducted to study whether subjects are capable of distinguishing the speaker from the listeners according to the gaze behaviors of the interlocutors alone. Our results show that eye contact provides a reliable cue for the identification of the speaker in three-party conversations.

ACM Classification Keywords
H.1.2. User/Machine Systems: Human information processing

Author Keywords
eye gaze; eye contact; face-to-face communication; multiparty conversation; human-human interaction; head gestures; eye-head coordination; perception of gaze; nonverbal behaviors.

INTRODUCTION
Occurring in the visual channel, nonverbal behaviors contribute to the flow of conversation including gaze, facial expressions, head rotations, gestures, etc. These behaviors are of importance in technology-mediated communication, such as videoconferencing and interactions with embodied conversational agents (ECAs) in virtual worlds. Indeed, the rendition of natural nonverbal behaviors is essential to increase the appreciation and the effectiveness of mediated interactions [7, 39, 11, 6, 14].

To improve the quality of mediated interaction, tremendous progresses have been made to synthesize nonverbal behaviors for virtual humans as the speaker or a listener in a conversation. Often, the synthesized speaker behaviors are assumed to be independent of listeners [5, 4, 15, 25, 24, 27, 9, 8, 10]. On the other hand, the behaviors of virtual listeners in some previous works [26, 30, 29] are typically driven by the detected motions of the user (i.e., the human speaker), which implicitly rely on an over-simplified hypothesis that the listener behaviors are submissive to those of the speaker.

It could be untrue that the behaviors of the speaker and listeners are independent of each other or that one is submissive to the other one. Indeed, the interactions between two interlocutors exhibit varying degrees and patterns of mutual influence along several aspects such as talking style/prosody, gestural behavior, engagement level, emotion, and many other types of user states [3], which implies that the speaker and listener behaviors interact with each other. In the works mentioned, such interactions have not yet been merged into the synthesized animations for virtual speakers and listeners. This could probably result from that it is also unclear whether merging such interactions can improve the quality of mediated interactions.

Additionally, the above works targeting two-party conversations on behavior synthesis could not be straightforwardly extended to multiparty conversations due to the obvious change of the spatial arrangement between the speaker and the listeners. Such a gap could probably result from that the occurrences of a specific interaction, namely eye contact in multiparty conversations, have been relatively understudied to date.

In light of the above gaps, an intriguing yet widely open research question is, whether the rendition of interactions of nonverbal behaviors can improve the quality of mediated interaction. As the first step, we focus on the specific interaction of eye contact and conduct a multifaceted study on whether the speaker can be effectively identified from the interlocutors in

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05 $15.00
DOI: http://dx.doi.org/10.1145/3025453.3025644

a three-party conversation based on eye contact, which could provide detailed and precise understanding of human-human interaction than the direct rendition of eye contact. This would also provide indisputable evidence that eye contact can offer informative feature to the synthesis of nonverbal behaviors.

Eye contact occurs when two interlocutors are looking at each other. It is determined by the gaze directions of the interlocutors. The gaze is manifested via the combination of head orientation and eye orientation [22, 23]. It can be determined by head orientation only if the eyes keep still at the center of the orbit. So, it is critical to investigate the difference between the contributions of head and of eyeball to the identification of the speaker from the interlocutors, which could facilitate researchers to work on the better synthesis of gaze, head and eyeball in various applications.

The seminal work by Rienks et al. [35] has reported that people can use knowledge of head motions in the azimuth plane to distinguish the speaker from the listeners but that head alone does not provide a sufficient cue for reliable identification of the speaker in a multiparty conversation. Their investigation is carried out through subjective judgments where human observers are invited to watch the video clips of four virtual humans in a meeting and to identify the speaking virtual human. The virtual humans are only rotating their heads in the azimuth plane by displaying pre-recorded human data; neither other nonverbal behaviors are rendered nor speech signals are provided. Moreover, in each video clip, only one speaker is speaking, without turn-taking, overlapping speech, and silence.

Our study relies on a collected three-party conversational behavior dataset. The simultaneously collected data include the azimuth angles of head and eyeball and speech signals. Such data enables us to get insights into eye contact occurrences in a conversation. The objective and subjective studies are both carried out to identify the speaker from the interlocutors in three-party conversations. In the objective study, a Markov process is trained to encode the differences of eye contact patterns according to the interlocutors' statuses (speaking or listening), and then it is used to distinguish the speaker from the listener(s) using eye contact information as the input. In the subjective study, participants are invited to judge who is speaking and who is/are listening by observing the gaze behaviors of virtual humans in a conversation without speech signals, where the virtual humans are animated by pre-recorded human movement data. Moreover, our work takes insight into the trend of the identification accuracy along with the increasing length of speaker turn in both objective and subjective studies.

To the best of our knowledge, the only previous work that is directly relevant to our work is the aforementioned work [35]. There are several major differences between our work and [35]:

1. Our study is exactly based on eye contact behavior, while technically speaking their work is not. Specifically, our study considers both the head orientation and eyeball behavior to estimate eye contact, but their work considered head orientation alone. Indeed, their work fundamentally relies on a key assumption that the participants' focus of attention can be well approximated by head orientation. However, our study proves this key assumption in a multiparty setting may not always be valid, and found the gaze behaviors estimated from the joint head orientation and eyeball movement data clearly outperform head orientations alone, in terms of speaker identification in a three-party setting.

2. The work [35] did not provide experimental evidence to support the trend of the identification accuracy along with the increasing length of speaker turn. By contrast, our objective analysis method concretely shows that the speaker identification accuracy of using eye contact as the cue is more than 70% when the length of speaker turn is longer than 2 seconds (see Figure 7). Also, our subjective study results (see Figure 11) show that the identification accuracy is more than 80% when the speaker turn is longer than 3 seconds; the accuracy can be lower only if the length of turn speaker is less than 1 second. This is a new finding that has never been reported previously.

3. The work [35] mainly used subjective studies for speaker identification. Statistical objective analysis on the recorded human motion data was not conducted. By contrast, our work performs both statistical objective analysis and subjective studies to look into this problem from a multifaceted perspective, which gives new insights on this research problem of speaker identification.

BACKGROUND

Speaker diarization is an emerging research topic for automatic conversation/meeting analysis. It aims at determining "who spoke when?" in audio or video recording, where turn-taking often occurs and multiple speakers could simultaneously speak (overlapping speech). To achieve the goal, besides speech features, visual features are often also used to improve the recognition accuracy [12, 16, 20, 32]. These visual features characterize only individual behaviors, while interactive functions of behaviors are neglected, e.g., eye contact.

None of the previous works on speaker diarization explicitly employs eye contact as a cue to improve the recognition accuracy. It probably attributes to the lacking of awareness of eye contact occurrence between the speaker and the listener(s). Comprehensively reviewing all the prior efforts on speaker diarization is beyond the scope of this paper. Readers of interest are referred to recent survey articles [1, 38].

Our work differs from speaker diarization. Our aim is to obtain insights on the precise understanding of human-human interaction by eye contact, according to the interlocutors' statuses (speaking or listening). Our work could offer researchers a new insight into the relation between interlocutor statuses and eye contact as well as provide a new useful clue to improve the state of the art in speaker diarization.

Predictions of turn-taking/keeping, the next speaker, and the lasting time between two speaker-turns can be carried out according to interactive gaze patterns, such as convergence, divergence, and eye contact (dyad-link), which was reported in some previous works [17, 18, 19, 21, 34]. Their positive

results indicate that interactive gaze patterns are closely related to the occurrences of turn-taking and turn-keeping. Moreover, their works suggest that humans are very skilled at encoding and decoding interactive gaze patterns for smoothing the flow of conversation.

Although gaze patterns used in these works occur in the phase of turn-taking/keeping, i.e., the last 1000 ms of an utterance and the follow-up pause (200 ms), rather than during the whole utterance, the successful use of interactive gaze patterns for identifying turn-taking and turn-keeping inspires us how to design our analysis.

Identifications of interactive gaze patterns (mentioned above) and of conversation regime (such as monologue and discussion) can be conducted according to head orientations and who is speaking, which is reported in previous works [33, 13]. Their validated results suggest that gaze patterns are correlated with the interlocutors' statuses (speaking or listening).

CONTRIBUTIONS

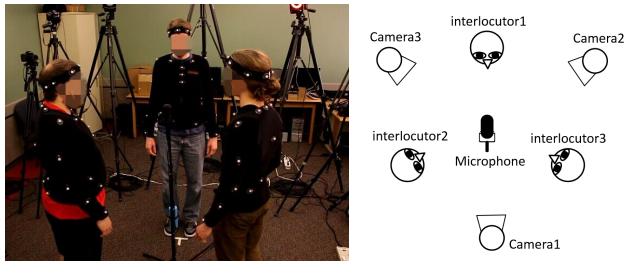
Our work contributes to the understanding of eye contact patterns and of head and eyeball movements in three-party conversations. It provides an important clue to extend the existing works on head and eyeball animation generation to multi-party conversations, and to bridge the gap between listener and speaker animations. In addition, this work provides a new clue to an emerging research topic of speaker diarization and to automatic conversation/meeting analysis (video editing, navigation, retrieval, or higher-level inference).

DATA ACQUISITION AND PROCESSING

In this work, we collected a multimodal conversational behavior dataset for three-party conversations, using the acquisition setup described below. In total, 2 three-party conversation datasets were captured, which involved 6 different participants. Each three-party conversation dataset consists of 6 sessions, each of which lasted about ten minutes. The participants took a break of several minutes between two successive sessions. The participants of one three-party conversation dataset are 3 males and those of the other are 3 females. The participants (ages are from 19 to 25) did not know each other before the experiment. No instructions were made into their conversations; they freely talked with each other on the topics of self-introduction, life, hobbies, favorite restaurants and movies, traveling plan, career plan, etc.

Acquisition setup. The datasets of three-party conversation were recorded from three interlocutors standing at the three vertices of an equilateral triangle with a distance of about 1 meter to other interlocutors, which is designed to eliminate any potential bias from the sitting position and group formation. The interlocutors were instructed to stick to their original locations as much as possible, but they were allowed to naturally move their torsos and heads, and change facial expressions during the conversation. A snapshot and the layout of the three-party conversation capture are shown in Figure 1.

The collected conversational behavior data contain body movement, eyeball movement, eyelid movement, and acoustic speech, all of which were acquired simultaneously at 30 frames



(a) (b)

Figure 1. The layout of the recorded three-party conversations. (a) A snapshot of the data acquisition process. (b) Illustration of the eye tracking data acquisition process.

Figure 2. Detection of eyeballs movement. The detected eye and pupils are rendered respectively by the green dashed lines and the solid red circles. As can be seen, the solid red circles almost cover the human pupils. The images from left to right display a detected gaze aversion from left to right.

per second. The body movement includes hand gesture, torso movement, and head movement. The body movement was recorded by a ten-cameras VICON optical motion capture system, where each interlocutor wore a mocap suit attached with optical markers. We downsampled the mocap data from the originally captured 120 frames/second to 30 frames/second. The torso and head movements were recorded as the sequences of 3-dimensional skeleton rotation angles. Moreover, acoustic speech was recorded by a wireless microphone positioned in the center of the interlocutors (see Figure 1). The speech transcription and its time information were obtained manually by a language expert.

To acquire the accurate 3D movements of the eyeballs and the eyelids of all the interlocutors, we employed the proven, hybrid mocap and HD video capture scheme proposed by Binh et al. [24]: specifically, the movement of the front face of each interlocutor was recorded by a Cannon HD camera (Figure 1-(b)). Then, with the aid of accurate 3D head movement at each frame that can be accurately estimated from several mocap markers on the head, a specially-designed tracking algorithm was used to offline process the video to track the movements of the eyeballs and the eyelids frame by frame [24]. When the eyelids were detected as being closed, the eyeballs movement could not be detected and it was inferred by smoothing the movement trajectory with linear interpolation. Figure 2 shows a few images of the acquired eyeball movements.

Gaze is approximated by adding the azimuth angles of the eyeballs, the head, and the torso. In our work, the rotational angle of the torso is always merged into that of the head; it is viewed as a part of the head rotational angle. So the head rotational angle in the remaining writing represents the combination of the azimuth angles of the head and the torso.

Considering the locations of the interlocutors at the three vertices of an equilateral triangle, we assume that an interlocutor looks at none of the other interlocutors if his/her eye gaze ori-

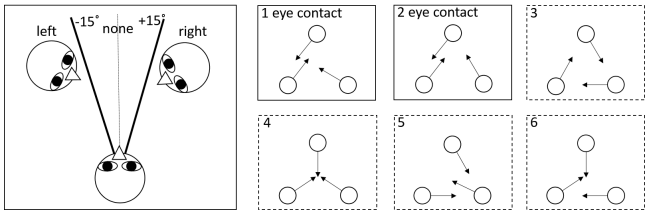


Figure 3. Illustrations of discrete gaze labels, of eye contact and of no eye contact. The left image shows that the gaze at each time frame, estimated from head and eyeball data or from only head data, is labeled by three discrete gaze labels: left, none and right. The gaze between 15° to the left and 15° to the right is labeled by *none*; the gaze beyond 15° to the left/right is labeled by *left* / *right*. In the right six images, the images with No.1 and No.2 show two cases of eye contact. The other images show four cases of none eye contact.

ents in the range between -15° (left) and +15° (right)[1]. Also we assume that he/she looks at another interlocutor to his/her left or right if his/her eye gaze orients beyond 15° to left/right. Based on these assumptions, the eye gaze of each interlocutor at each frame can be mapped to one of the three discrete labels: looking at none of the other interlocutors (*none*), looking at the right interlocutor (*right*), and looking at the left interlocutor ($left$), which is illustrated in the left panel of Figure 3.

Eye contact is detected at each time frame by judging whether any two of three interlocutors are looking at each other, which is determined from their gaze labels (*none*, *right* and $left$). The right panel of Figure 3 shows two cases with eye contact and four cased without eye contact. To describe the occurrence of eye contact between them at each time frame, a 3-dimensional feature vector, s^i, is defined and it has four possible values as follows.

1. $s^1 = [0,0,0]$: eye contact does not occur.
2. $s^2 = [1,1,0]$: eye contact occurs between the 1st and the 2nd interlocutors.
3. $s^3 = [0,1,1]$: eye contact occurs between the 2nd and the 3rd interlocutors.
4. $s^4 = [1,0,1]$: eye contact occurs between the 1st and the 3rd interlocutors.

s^i consists of three elements, which respectively represent whether the three interlocutors involve eye contact or not. The element can be 0 or 1. 0 stands for no involvement of eye contact; 1 stands for the involvement of eye contact. Furthermore, $S^k = [s_1, ..., s_t, ..., s_T]$, a sequence of s^i, is used to describe the occurrences of eye contact along with time frame. k in S^k stands for the index of motion episode with one specific speaker speaking, which is clipped from the whole recorded motion stream, according to speech information. The set of S^k is noted by $\{S^k\}$, $k = 1,2,...,K$, where K is the total number of motion episodes. In our dataset, 2486 motion episodes are collected (K=2486) with the duration ranging from 0.5 second to 20.3 seconds (averaged duration is 2.27 seconds and SD = 2.40). Note that those motion episodes with length less than 0.5 second are screened out as done in Rienks et al. [35].

Note that although the peripheral vision of an interlocutor may also include other interlocutor(s) besides the primary eye-contact target, at a particular time instance an interlocutor can at most gaze at one other interlocutor as the focus of his/her attention. Therefore, despite certain loss of information, our study treats eye contact as a binary variable. Ignoring the peripheral vision could still retain the most prominent information of the interlocutors' focus of attention for our study.

EXPERIMENT: IDENTIFYING THE SPEAKER

In our work, both subjective judgement and objective analysis are conducted. In the objective analysis, eye contact (a specific interactive behavior) rather than individual gaze (or head orientation) is taken into account as input features. It is based on the assumption that eye contact could provide more useful cues than individual gazes as it displays interactive behaviors between interlocutors. Moreover, in the subjective judgment, the gaze azimuth angles are used to animate virtual humans in a conversation, where whether eye contact occurs or not is not explicitly indicated.

Our first set of hypotheses are:

- **H1-obj**: objective eye contact data can be used to effectively identify the speaker from the interlocutors in three-party conversations (objective analysis).
- **H1-sub**: humans are capable of using knowledge of gaze patterns to distinguish the speaker from the listeners in three-party conversations (subjective judgment).

The investigation of H1-obj and H1-sub could offer new insight into the understanding of interactive gaze patterns among interlocutors in three-party conversations.

Furthermore, eye contact and gaze are estimated by combining the eyeball and head rotation angles. Such a combination could provide more precise information about the attention of interlocutors than head rotation angles alone. Our second set of hypotheses are:

- **H2-obj**: objective analysis can identify the speaker with a higher accuracy from the joint head and eyeball movement data than the head movement data alone.
- **H2-sub**: subjective judgment can identify the speaker with a higher accuracy from the joint head and eyeball movement data than the head movement data alone.

The investigation of H2-obj and H2-sub allows us to understand the contribution of eyeballs to face-to-face communication in three-party conversations.

The objective analysis and subjective judgment aim at determining who is speaking according to the head and eyeball azimuth angles in a speaker turn, which could last between 0.5 second and 20.3 seconds in our dataset. Considering that long speaker turns could provide more information for the speaker identification than short ones, we formulate our third set of hypotheses as follows:

[1]A common strategy employed in graphics literature is to define a cutoff displacement of 10°-15° above which targets are acquired by a gaze shift [36][28][31]. In our work, 15° is used to distinguish the gaze target.

- **H3-obj**: objective analysis can obtain a higher speaker identification accuracy from long speaker turns than from those short ones.

- **H3-sub**: subjective judgment can obtain a higher speaker identification accuracy from long speaker turns than from those short ones.

In the following writing, we will describe the conducted objective analysis and subjective judgment including methodology, results, and discussion.

Objective Analysis

To objectively identify the speaker, a Markov process can be built based on $\{S^k\}$. In each experiment, the used set of S^k is splitted into a training set and a test set. Two pre-processing steps are performed to the training set before a Markov process is trained. Note that such pre-processing steps are not applied to the test set.

Building Markov Process

Assuming S^k is one of the motion episodes in the training set, the first pre-processing is to make the first element of s_t, t=1,...,T, in S^k always correspond to the known speaker by switching the 1st interlocutor and the known index of the speaker in s_t. For example, if the 1st, the 2nd and the 3rd elements in s_t respectively correspond to one listener, the other listener, and the speaker, the 3rd and the 1st elements are switched. If the 1st element originally corresponds to the speaker, the order of the elements are kept without switching. After this manipulation, while the 1st element of s_t corresponds to the speaker, the 2nd and the 3rd elements correspond to the listeners.

Afterwards, the second pre-processing is to create a new set by switching the 2nd and the 3rd elements of all the $\{s_t\}$ in the original training set, which corresponds to two listeners (see the first pre-processing above). Then, we combine the new created set with the original training set to obtain the new training set that is used to build the Markov process. In our recorded data, two listeners stand symmetrically around the speaker. Therefore, this new training set is designed to eliminate any potential bias from the right or the left listener.

The built Markov process, λ, consists of four states, each of which corresponds to one feature vector, s^i, where $i = 1,2,3$ or 4. The transition probabilities are calculated from the training set, by counting and normalizing the number of transitions between the feature vectors.

The built Markov process is expected to capture the transition probabilities of four eye contact patterns (s^1, s^2, s^3, and s^4), viewing the speaker as the first interlocutor. How to use the Markov process to identify the speaker is described below.

Using Markov Process

The aim of speaker identification is to infer the speaker index from S^k in the test set, using λ. At the identification step, the speaker index is unknown in the test set; it can be the 1st, the 2nd, or the 3rd. Given S^k, a probability can be inferred from the transition probabilities at all the time steps, and it quantifies the probability of S^k from the Markov process. The probability calculation is detailed below.

Since the Markov process is trained on the set of $\{S^k\}$ with the speaker being the first interlocutor, we expect that a test speaker turn, $S^{k_1} = [s_1^{k_1}, ..., s_{T_{k1}}^{k_1}]$, results in a higher probability when the first interlocutor is the speaker than when he/she is a listener. Based on this hypothesis, we alter the interlocutors' indexes twice. The first is to switch the 1st and the 2nd elements of $\{s_t^{k_1}\}_{t=1}^{T_{k1}}$ and the altered speaker turn is denoted by $S^{k_2} = [s_1^{k_2}, ..., s_{T_{k1}}^{k_2}]$; the second is to switch the 1st and the 3rd elements of $\{s_t^{k_1}\}_{t=1}^{T_{k1}}$ and the altered speaker turn is denoted by $S^{k_3} = [s_1^{k_3}, ..., s_{T_{k1}}^{k_3}]$. We then applied λ to S^{k_1} as well as S^{k_2} and S^{k_3} respectively to obtain their probabilities. To this end, S^{k_i} with the maximum probability is selected, and the i is selected as the speaker index.

The probability of S^{k_i} is calculated as follows.

$$p(S^{k_i}|\lambda) = \prod_{t=1}^{T_{k1}-1} \frac{P(s_{t+1}^{k_i}|s_t^{k_i}, \lambda)}{\sum_{1\le j\le 3} P(s_{t+1}^{k_j}|s_t^{k_j}, \lambda)} \tag{1}$$

where $p(S^{k_i}|\lambda)$ stands for the probability of S^{k_i}, given the built Markov process; $P(s_{t+1}^{k_i}|s_t^{k_i}, \lambda)$ is the transition probability of feature vectors from $s_t^{k_i}$ to $s_{t+1}^{k_i}$, which also applies to $P(s_{t+1}^{k_j}|s_t^{k_j}, \lambda)$. To avoid the problem of numeric overflow, the operation of summation is to normalize the probabilities so that the probabilities of three variations are summed to be 1 at each time step, as follows.

$$\sum_{1\le i\le 3} \frac{P(s_{t+1}^{k_i}|s_t^{k_i}, \lambda)}{\sum_{1\le j\le 3} P(s_{t+1}^{k_j}|s_t^{k_j}, \lambda)} = 1 \tag{2}$$

The above objective analysis is applied to validate the hypotheses H1-obj, H2-obj, and H3-obj. According to the hypotheses, more details as well as the results will be described below.

Objective Analysis Results

Test of the hypotheses H1-obj and H2-obj. Two types of Markov processes are built based on $\{S^k\}$. The first one, denoted by $\{S^k\}_{he}$, is built from the eye contact data that is estimated from the joint azimuth angles of the head and the eyeballs; the second one, denoted by $\{S^k\}_h$, is done from the eye contact data that is estimated from the azimuth angles of the head alone. In each experiment, $\{s^k\}_{he}$ and $\{s^k\}_h$ are respectively randomly splitted into a training set (80%) and a test set (20%). The training sets from $\{s^k\}_{he}$ and $\{s^k\}_h$ are used to respectively build two Markov processes, noted by λ_{he} and λ_h. Then, the test sets from $\{s^k\}_{he}$ and $\{s^k\}_h$ are respectively applied to λ_{he} and λ_h. We ran the above *repeated random sub-sampling validation* experiments 20 times for $\{s^k\}_{he}$ and $\{s^k\}_h$, respectively. To this end, by comparing the recognition result with the ground-truth, we can obtain a recognition accuracy rate in each experiment. The averaged experiment results are reported in Figure 4 and Figure 7.

To test the hypothesis H1-obj, we compared our experiment results with the baseline (33.3%) that is obtained by random

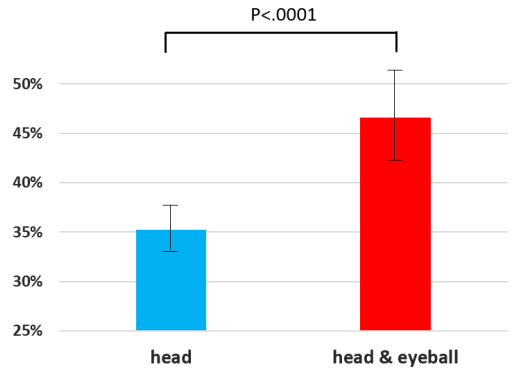


Figure 4. Average results of speaker identification in our objective analysis. The vertical bars visualize the standard deviations.

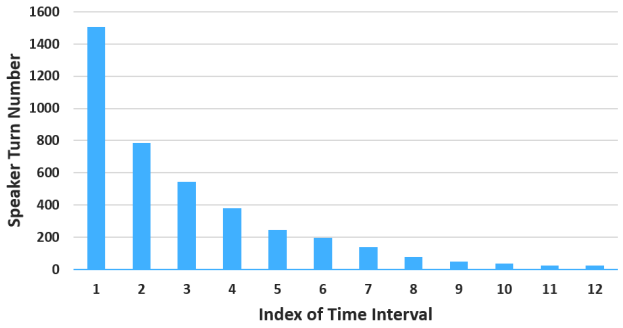


Figure 5. The number of speaker turns in each Interval. The intervals are defined by a 2-seconds moving window and a step size of 1 second. As the exceptions, the 1st and 12th intervals contain the speaker turns whose lengths are in the range of $[0.5, 2]$ and $[11, 20.3]$, respectively. Each speaker turn consists of many frames. The total frame number of all the speaker turns in each interval can be found in Figure 6.

selection (one out of three interlocutors). We conducted One-Sided t-Test on the results of 20 validation experiments. For $\{s^k\}_h$, the recognition accuracy (M=0.376, SD=0.009) is significantly greater than the baseline: t(19)=20.7967, p<0.0001. For $\{s^k\}_{he}$, the recognition accuracy (M=0.4664, SD=0.0165) is also significantly greater than the baseline: t(19)=35.2144, p<0.0001. Therefore, the hypothesis H1-obj is accepted (assuming the significance level = 0.05).

To test the hypothesis H2-obj, we conducted a Paired-Sample t-Test to compare the 20 validation experiment results between λ_{he} and λ_h. The comparison results indicate that the joint head and eyeball movement data (M=0.4664, SD=0.0165) result in a significantly greater recognition accuracy than the head movement data alone (M=0.376, SD=0.009): t(19)=22.8802,p<0.0001. The hypothesis H2-obj is accepted.

Test of the hypothesis H3-obj. To take into account the length of speaker turn, the length range is splitted into overlapping intervals with a 2-seconds moving window and a step size of 1 second. Each speaker turn is classified into two successive intervals. For example, one speaker turn with the length of 2.3 seconds is grouped into two successive intervals in the ranges of $[1, 3]$ and $[2, 4]$. The number of speaker turns in each

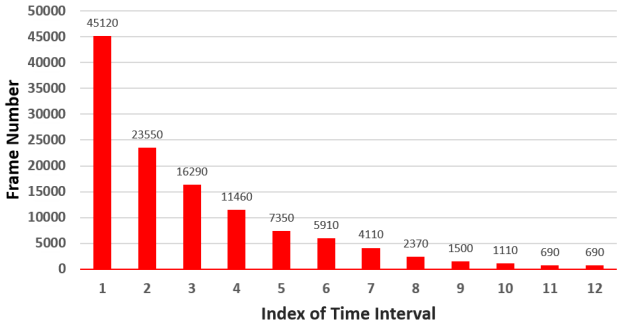


Figure 6. The frame number of all the speaker turns in each Interval. The intervals are defined by a 2-seconds moving window and a step size of 1 second. As the exceptions, the 1st and 12th intervals contain the speaker turns whose lengths are in the range of $[0.5, 2]$ and $[11, 20.3]$, respectively. The number of speaker turns in each interval can be found in Figure 5.

interval is illustrated in Figure 5. In fact, the 1st interval in the range of $(0, 2]$ only contains those speaker turns whose lengths are in the range of $[0.5, 2]$, since we discarded those speaker turns with the length less than 0.5 second in our dataset (refer to the aforementioned data acquisition and processing section). As can be seen from Figure 5, the number of speaker turns decreases with the increase of the length of speaker turn. As such, the $12th$ interval in the figure contains all the speaker turns with the length more than 11 seconds (i.e., in the range of $[11, 20.3]$). Each speaker turn is characterized by a sequence of frames. The frame number of all the speaker turns in each time interval is shown in Figure 6.

The speaker turns from a specific interval are used to calculate the identification accuracy for this interval. That is, each interval is associated with an identification accuracy. To calculate the recognition accuracy for each interval, the used set of S^k only involves the speaker turns in a specific interval. In each experiment, 80% of the speaker turns are used as the training samples and 20% as the test samples. To this end, two estimates of $\{S^k\}$, $\{s^k\}_{he}$ from the joint head and eyeball movement data and $\{s^k\}_h$ from the head movement data alone are investigated. Figure 7 depicts the result of each time interval.

The red points in Figure 7 represent the recognition results from the joint head and eyeball movement data. As illustrated in this figure, it rapidly increases between the $1st$ and the $4th$ intervals; it stays steady between the $4th$ and the $7th$ intervals; and it increases with obvious fluctuations between the $7th$ and the $12th$ intervals. These observations suggest its overall increasing trend despite its fluctuations at some places. However, it does not monotonically increase. More investigations are needed to look into those fluctuations as the future work. Similar observations can also be seen from the blue curve representing the recognition results based on the head movement data alone, although the increasing trend of the blue curve is less obvious than that of the red one.

In addition, the red points in Figure 7 are consistently higher than the blue points. The joint head and eyeball movement data results in a higher recognition accuracy in every interval

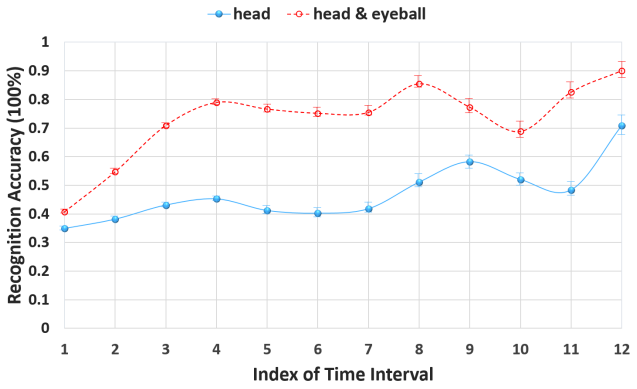


Figure 7. The objective speaker identification accuracy as a function of the length of speaker turn. The red dashed curve shows the results from the joint head and eyeball movement data while the blue solid curve shows the results from the head movement data alone. The plotted results are obtained by averaging the outcomes of 20 repeated random subsampling validation experiments. The standard deviations are plotted as vertical bars.

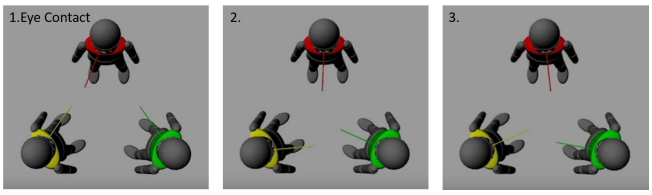


Figure 8. Several snapshots of rendered virtual character animation clips in our study. The colored lines represent the gaze orientations of the virtual characters. The 1st panel shows the occurrence of eye contact; the other two panels show two cases of no eye contact.

than the head movement data alone, which is in line with the validated hypothesis H2-obj. Moreover, we observe that, the speaker identification accuracy from the joint head and eyeball movement data or from the head movement data alone is clearly higher than the baseline (33.3%) in every interval, which is also in line with the validated hypothesis H1-obj.

In Figure 7, the accuracy from the joint head and eyeball data is about 40% in the 1*st* time interval, and the accuracies in other time intervals (i.e., since the 3rd interval) are more than 70%. However, as shown in Figure 5, the number of the speaker turns in the 1*st* time interval is much higher than those in the other time intervals. That is why the averaged accuracy (counting all the intervals) from the joint head and eyeball data in Figure 4 is about 46%.

Subjective Judgment

Subjective judgments are obtained through a user study, where participants are invited to watch episode animation clips displaying three rendered virtual characters (interlocutors) in a conversation. Each animation clip contains a single speaker turn where only a virtual character is speaking and the other two are listening without turn-taking and overlapping speech. The animations are rendered from the top view of the virtual characters (refer to Figure 8). The virtual characters are animated only by the gaze azimuth data and speech signals are not provided. Gaze orientation is represented by a solid line, which allows participants to clearly observe the gazes of the virtual characters at each time instant. In the animations, torsos and gaze lines of the three virtual characters are colored by red, green, and yellow, respectively. The occurrences of eye contact are not explicitly indicated in the animation clips. The animation clips are played at a speed of 30 frames per second, as the originally recorded human movement data.

To test the hypothesis H1-sub, gazes are estimated by the joint head and eyeball movement data. The hypothesis H1-sub is accepted if the subjective identification accuracy is statistically higher than the baseline (33.3%). To test the hypothesis H2-sub, gazes are estimated not only by the joint head and eyeball movement data but also by the head movement data alone. The hypothesis H2-sub is accepted if the subjective identification accuracy on the animations driven by the joint head and eyeball movement data is statistically higher than that by the head movement data alone. To test the hypothesis H3-sub, the duration of speaker turn need to be taken into account. The hypothesis H3-sub is accepted if the subjective recognition accuracy increases along with the length of speaker turn.

Protocol: Participants are asked to provide demographic information, including age, gender, education level, occupation, etc., before the experiment; then, they are invited to view animation clips of three virtual characters in a conversation. Their task is to choose the speaker in each clip. We describe the elements of the protocol in this user study as follows. (1) Participants: in total, there were 34 participants consisting of 22 males and 12 females with the age ranging from 21 to 70 (M=30.06 years, SD=9.9 years). (2) Stimuli: the range of speaker turn length was splitted into 5 non-overlapping intervals: 0.5 to 3 seconds, 3 to 6 seconds, 6 to 9 seconds, 9 to 12 seconds, and 12 to 20.3 seconds. 5 speaker turns were randomly selected from each interval. To this end, a total of 25 speaker turns were chosen (5 intervals x 5 samples per interval). The shortest selected sample lasts 1.1 seconds, and the longest one lasts 13.6 seconds. Furthermore, 2 different versions (conditions) of the virtual character animations were created for each speaker turn:

- **Condition 1**: gaze is estimated based on head orientation alone.
- **Condition 2**: gaze is estimated based on the joint head and eyeball orientations.

In total, 50 animation clips (without speech) were generated, which consists of 25 pair-wise animation clips with Conditions 1 and 2.

Procedure: The 50 animation clips are presented to each participant in a random order (refer to Figure 9). When participants watch the animation clips, they are explicitly informed that they can watch these animations as many times as they want; that only one interlocutor is speaking without turn-taking, and that the lines represent the gaze orientations of the virtual characters. Meanwhile, they are not explicitly informed whether the gaze orientation is estimated by the joint head and eyeball movement data or by the head movement data alone. After watching each animation clip, the participants are asked to select which character (red, green, or yellow) is the speaker in the animation, described below.

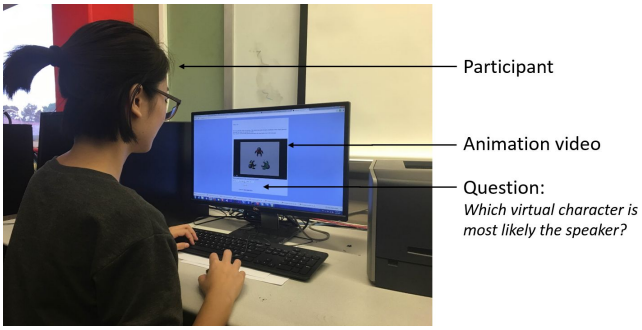


Figure 9. A user study snapshot of a participant

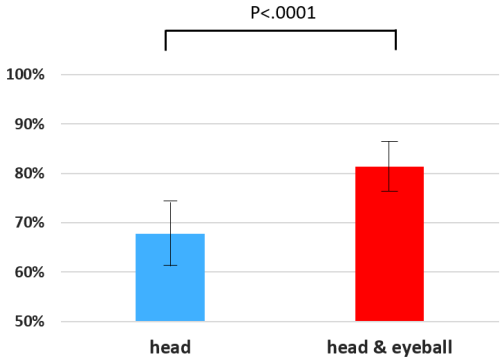


Figure 10. Overall subjective speaker identification accuracies (blue for condition 1 and red for condition 2). Standard deviations are plotted as the vertical bars.

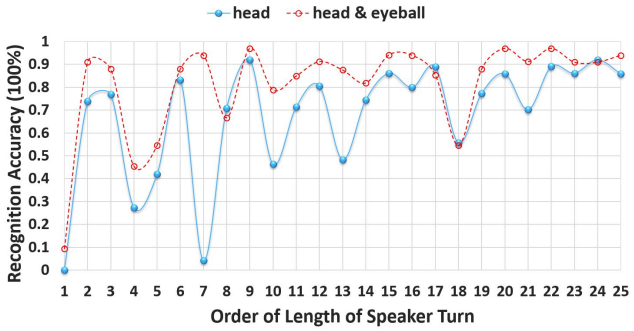


Figure 11. Subjective speaker identification accuracy as a function of speaker turn length. The red hollow circles show the results with Condition 2 (joint head and eyeball motion data) while the blue solid circles show the results with Condition 1 (only head motion data). In 21 out of 25 pairwise comparisons, Condition 2 outperforms Condition 1, while Condition 1 outperforms Condition 2 in 4 pairwise comparisons, whose indexes are 8, 17, 18 and 24.

- Which virtual character is most likely the speaker?

Results: We report two types of recognition accuracies, whose difference is whether the length of speaker turn is viewed as a dependent variable. The length of speaker turn is not taken as a dependent variable in the first type while it is taken in the second type. The condition mentioned above is always taken as a dependent variable.

Considering the condition as a dependent variable, the first type contains two recognition accuracies: one is calculated based on the subjective judgment results from those 25 animation clips with Condition 1, and the other one is calculated based on the subjective judgement results from those 25 animation clips with Condition 2. The results are reported in Figure 10.

The second type is to calculate the recognition accuracy for each animation clip. The results are reported in Figure 11, where the blue and red points represent the results with Condition 1 and Condition 2, respectively. In this figure, the vertical axis represents the recognition accuracy, and the horizontal axis represents the index of an animation clip. Note that the animation clips in this figure are ordered by their lengths (i.e., the length of speaker turn) in an increasing order. The red and blue points positioned with the same values in the horizontal axis are associated with the two pair-wise animations (corresponding to Conditions 1 and 2, respectively).

Test of the hypotheses H1-sub and H2-sub. To test the hypothesis H1-sub, the first type of recognition accuracy is compared with the baseline (33.3%). The results are shown in Figure 10. One-sided t-Test is conducted on the subjective judgment results from the 25 animation clips with Condition 1 and the other 25 clips with Condition 2. According to the results with Condition 1, the recognition accuracy (M=67.7%, SD=0.25) is statistically higher than the baseline: $t(24)=6.92$, $p<.0001$. Also, according to the results with Condition 2, the recognition accuracy (M=81.4%, SD=0.21) is statistically higher than the baseline: $t(24)=11.75$, $p<.0001$. The hypothesis H1-sub is accepted not only for Condition 1 but also for Condition 2.

To test the hypothesis H2-sub, a paired-sample t-Test is conducted to compare the subjective judgment results with Conditions 1 and 2. The comparison results show that the results with Condition 2 (M=81.4%, SD=0.21) lead to statistically higher recognition accuracies than those with Condition 1 (M=67.7%, SD=0.25): $t(24)=3.65$, $p<.01$. Therefore, the hypothesis H2-sub is accepted.

Test of the hypothesis H3-sub. To test the hypothesis H3-sub, the second type of recognition accuracy is taken into account, which is illustrated in Figure 11. As shown in this figure, the red and the blue curves do not show any obvious (increasing or decreasing) trend along with the length of speaker turn, which does not provide sufficient evidence to statistically test the hypothesis H3-sub.

Additionally, as observed in Figure 11, Condition 2 outperforms Condition 1 in 21 pair-wise comparisons while Condition 1 outperforms Condition 2 in 4 pair-wise comparisons ($8th$, $17th$, $18th$ and $24th$). In other words, Condition 1 outperforms Condition 2 with the probability of $16\%(=4/25)$. This implies that the hypothesis H3-sub cannot be rigorously validated for each pair-wise comparison, although in our experiments it holds true with the probability of $84\%(=21/25)$. On the other hand, the probability of the hypothesis H3-sub holding true, 84%, is substantially larger than the probability of the hypothesis H3-sub holding false, 16%, which is in line with the validated hypothesis H2-sub.

Furthermore, the recognition accuracies shown by the red points in Figure 11 are higher than the baseline (33.3%) except for the shortest (1st) speaker turn that lasts 1.1 seconds, while most of the blue points are higher than the baseline except for the shortest speaker turn (1.1 seconds) and the $7th$ one (6.2 seconds). These observations provide empirical evidence that in general the hypothesis H1-sub (either condition 1 or condition 2) holds true, in particular, when the length of speaker turn becomes large.

DISCUSSION

In this section, we discuss the experimental results from a multifaceted perspective, including both the objective analysis and the subjective judgment, along with the formulated hypotheses.

Discussion on the Hypotheses H1-obj and H1-sub

Both the hypotheses H1-obj and H1-sub are validated by the joint head and eyeball azimuth angles (Condition 2). This suggests that the joint head and eyeball movement data provides a reliable cue to identify the speaker from the interlocutors in a three-party conversation, which can be not only captured by the proposed objective analysis method but also perceived by human subjects.

The statistical validation of the hypothesis H1-obj suggests that the interactive behavior of eye contact plays an important role in speaker identification. This also serves an empirical foundation to properly merge eye contact into synthesized head and eyeball animations in order to build more socially engaging, realistic conversational characters in various applications.

Rienks et al. [35] reported that humans use the knowledge of gaze patterns to distinguish the speaker from listeners, which is in line with the successful validation of the hypothesis H1-sub in our work. Although eye contact is not explicitly indicated in the used animation clips, the validation of the hypothesis H1-sub suggests that human viewers could be intrinsically skilled at decoding the occurrences of eye contact or other interactive gaze patterns (e.g., convergence or divergence).

Additionally, Rienks et al. [35] reported that gaze estimated by the head azimuth angle alone does not provide a sufficient cue for reliable speaker identification, which is not supported by the validated hypothesis H1-sub in our work. This difference could result from three factors. First, the video clips used in [35] record virtual humans from the side view, which degrades the perceptual judgment on the gaze orientations for participants, while the animation clips in our work are created from the top view, which allows the participants to faithfully perceive and observe the gaze orientations. Second, from the participants' perspective, the subtle rotation angle of gaze could be easily overlooked. To overcome this problem, a gaze line is used in our work for highlighting the gaze orientation with even subtle angle change. Third, more importantly, the eyeball rotation angle is not considered in their work, while it is used in our work to estimate the gaze orientation. This factor is also supported by the statistically validated hypotheses H2-obj and H2-sub.

Discussion on the hypotheses H2-obj and H2-sub

The statistical validation of the hypotheses H2-obj and H2-sub suggests that the joint head and eyeball movement data provides a more reliable cue to identify the speaker from the interlocutors than the head movement data alone. In our study, the contribution of eyeball movement to the speaker identification can be well captured by our objective analysis method as well as subjective perception study experiment.

In the validation of the hypothesis H2-obj, a significant gap between Conditions 1 and 2 indicates that the eyeball movement is critical to improve the recognition accuracy in the objective analysis process. Not surprisingly, in the validation of the hypothesis H2-sub, a similar significant gap exists between Conditions 1 and 2. This suggests that the role of eyeball movement in three-party conversations can be directly perceived by human observers. With being said, we also observe that the head movement data alone can provide a reasonable cue for speaker identification. This observation suggests that human viewers are sensible to head movement for conversation coordination and are skilled in decoding information from the head movement.

The statistical valiation of the hypotheses H2-obj and H2-sub also provide convincing evidence that the eyeball movement is crucial to accurately estimate human gaze in a conversation, which contradicts to the previous finding that gaze can be well approximated by head orientations alone [37, 2].

Discussion on the hypotheses H3-obj and H3-sub

The hypothesis H3-obj is partially true in the range from 1 to 4 seconds; alternatively, fluctuations are observed in the range from 6 to 12 seconds. Moreover, the overall recognition accuracy in the range from 6 to 12 seconds is higher than that in the range from 1 to 4 seconds. Coincidentally, such observations were also reported in Rienks et al. [35]. On one hand, the observed overall increasing trend can be explained by more information provided in a longer speaker turn. On the other hand, the fluctuations observed in long intervals could be explained as follows: the information complexity increases with the increased length of speaker turn so that the proposed Markov process falls short of well capturing the increasing information complexity; therefore, its performance has fluctations in this range.

Although the hypothesis H3-sub is not statistically validated by the subjective judgments in our work, we observe that the averaged recognition accuracy (81%) is much higher than baseline (33.3%). This suggests that humans are intrinsically skilled at identifying the speaker so that the identification accuracy is independent of the length of speaker turn, to a certain extent.

CONCLUSIONS

In this work, we have conducted a multifaceted study to understand whether the speaker can be rigorously identified from the interlocutors involved in three-party conversations based on their gaze/head orientations. Based on recorded three-party conversational behavior datasets, a Markov process based statistical framework is proposed to model the sequence of eye contact between the human interlocutors, and further used to

identify the speaker. Meanwhile, a user study is conducted to study this question using virtual characters in a conversation, animated by pre-recorded human movement data. Then, participants are invited to select the speaking virtual character by watching the animation clips of virtual three-party conversations.

Our results show that the proposed objective analysis method and humans are both capable of identifying the speaker and that interactive eye contact behaviors provide a rigorious cue for reliable speaker identification. Moreover, the gaze behaviors estimated from the joint head and eyeball movement data clearly outperforms those from head movement data alone. Additionally, the performance of the objective analysis depends on the length of speaker turn while human judgment is less dependent on it. Finally, our work indicates that the eyeball movement is essential to the accurate estimation of the gaze.

Some limitations exist in our current work:

1. This work relies on our definition of eye contact: eye contact is determined by the left-right angles of head and eyeball while the up-down angles are ignored. For some special cases such as interlocutors have very different heights, only the left-right angles could be insufficient to accurately estimate the occurrences of eye contact.

2. Eye contact used in the objective analysis is only one of the gaze interactive behaviors in multiparty conversations. The previous works [33, 13] have defined a few gaze interactive behaviors such as convergence and divergence. These interactive gaze behaviors are not considered in our current work.

3. The gaze lines used in the subjective judgment study play an important role in highlighting the gaze orientations. However, such visual representations may exaggerate the humans' perceptual capabilities on the conversational gaze.

4. The used datasets only contain three interlocutors standing at the three vertices of an equilateral triangle in a strict laboratory setting. It is still unclear whether our results can be rigorously generalized to other spatial arrangements of the interlocutors, to multiparty conversations with more than three interlocutors, or to coarse-grained gaze estimation. Furthermore, our work has not investigated the potential biases from individuals (e.g., gender, social status, emotion, etc.) and conversational topics.

5. Our work does not provide knowledge about what kind of gaze patterns are more reliable recognizer of speaker role and also overlooks the confidence of participants recognizing the speaker in the subjective experiments.

6. The participants in our subjective study acted as side observers instead of immersive interlocutors.

We will leave the tackling of the above limitations as the future work. Additionally, we will work on emerging the rendition of eye contact in the animation synthesis for virtual speaking and listening characters.

ACKNOWLEDGMENTS

This research is in part supported by NSF IIS-1524782 and National Natural Science Foundation of China Grant (No. 61328204). We would like to thank Binh Huy Le, Li Wei, Mingxuan Luo, Yaser Karbaschi, and Nafiseh Mehdipour for helping with the acquisition and processing of the used human motion dataset. We also wish to thank the anonymous CHI reviewers for their constructive comments.

REFERENCES

1. X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. 2012. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 2 (2012), 356–370.

2. A. C. Beall, J. N. Bailenson, J. Loomis, J. Blascovich, and C. S. Rex. 2003. Non-Zero-Sum Gaze in Immersive Virtual Environments. In *International Conference on Human-Computer Interaction*. 1108–1112.

3. J. K Burgoon, L. A. Stern, and L. Dillman. 1995. *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press.

4. C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan. 2007. Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. *IEEE Trans. on Audio, Speech & Language Processing* 15, 3 (2007), 1075–1086.

5. C. Busso, Z. Deng, U. Neumann, and S. Narayanan. 2005. Natural head motion synthesis driven by acoustic prosodic features. *JVCA* 16, 3-4 (2005), 283–290.

6. J. Cassell and K. R. Thórisson. 1999. The Power of a Nod and a Glance: Envelope Vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence* 13, 4 & 5 (1999), 519–538.

7. R. Alex Colburn, Michael F. Cohen, and Steven M. Drucker. 2000. *The Role of Eye Gaze in Avatar Mediated Conversational Interfaces*. Technical Report MSR-TR-2000-81. Microsoft Research. 9 pages.

8. Y. Ding, C. Pelachaud, and T. Artières. 2013. Modeling Multimodal Behaviors from Speech Prosody. In *International Conference on Intelligent Virtual Agents*. 217–228.

9. Y Ding, M Radenen, T Artières, and C Pelachaud. 2012. Eyebrow Motion Synthesis Driven by Speech. In *Workshop Affect, Compagnon Artificiel, Interaction (WACAI)*. 103–110.

10. Y. Ding, M. Radenen, T. Artières, and C. Pelachaud. 2013. Speech-driven Eyebrow Motion Synthesis With Contextual Markovian Models. In *ICASSP*. 3756–3760.

11. M. Garau, M. Slater, S. Bee, and M. A. Sasse. 2001. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 309–316.

12. D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan. 2007. Audiovisual probabilistic tracking of multiple

speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 2 (2007), 601–616.

13. S. Gorga and K. Otsuka. 2010. Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *ICMI-MLMI*. 54.
14. D.K.J. Heylen, I. van Es, A. Nijholt, and Dr. E.M.A.G. van Dijk. 2002. Experimenting with the Gaze of a Conversational Agent. In *Proceedings International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. 93–100.
15. G. Hofer and H. Shimodaira. 2007. Automatic Head Motion Prediction from Speech Data. In *Proc. Interspeech*. 722–725.
16. H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez. 2008. Associating audio-visual activity cues in a dominance estimation framework. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 1–6.
17. R. Ishii, S. Kumano, and K. Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *ICASSP*. 2319–2323.
18. R. Ishii, K. Otsuka, S. Kumano, and J. Yamato. 2014. Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings. In *ICASSP*. 694–698.
19. R. Ishii, K. Otsuka, S. Kumano, and J. Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Trans. Interact. Intell. Syst.* 6, 1 (2016), 4:1–4:31.
20. K. Ishizuka, S. Araki, K. Otsuka, T. Nakatani, and M. Fujimoto. 2009. A Speaker Diarization Method Based on the Probabilistic Fusion of Audio-visual Location Information. In *ICMI-MLMI*. 55–62.
21. T. Kawahara, T. Iwatate, and K. Takanashi. 2012. Prediction of Turn-Taking by Combining Prosodic and Eye-Gaze Information in Poster Conversations. In *INTERSPEECH*. 727–730.
22. C. L. Kleinke. 1986. Gaze and eye contact: A research review. *Psychological Bulletin* 100, 1 (1986), 78–100.
23. B. J. Lance and S. C. Marsella. 2008. A Model of Gaze for the Purpose of Emotional Expression in Virtual Embodied Agents. In *AAMAS*. 199–206.
24. B. H. Le, X. Ma, and Z. Deng. 2012. Live Speech Driven Head-and-Eye Motion Generators. *IEEE Transactions on Visualization and Computer Graphics* 18, 11 (2012), 1902–1914.
25. X. Ma and Z. Deng. 2009. Natural Eye Motion Synthesis by Modeling Gaze-Head Coupling. In *IEEE Virtual Reality Conference*. 143–150.
26. R. Maatman, J. Gratch, and S. Marsella. 2005. Natural behavior of a listening agent. In *International Workshop on IVA*. 25–36.
27. S. Mariooryad and C. Busso. 2012. Generating Human-Like Behaviors Using Joint, Speech-Driven Models for Conversational Agents. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2329–2340.
28. S. Masuko and J. Hoshino. 2007. Head-eye Animation Corresponding to a Conversation for CG Characters. *Comput. Graph. Forum* 26 (2007), 303–312.
29. L. Morency, I. de Kok, and J. Gratch. 2008. Predicting Listener Backchannels: A Probabilistic Multimodal Approach. In *International Conference on Intelligent Virtual Agents*. 176–190.
30. L. Morency, I. de Kok, and J. Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 70–84.
31. A. Normoyle, J. B. Badler, T. Fan, N. I. Badler, V. J. Cassol, and S. R. Musse. 2013. Evaluating Perceived Trust from Procedurally Animated Gaze. In *Motion in Games*. 141–148.
32. A. Noulas, G. Englebienne, and B. J. A. Krose. 2012. Multimodal Speaker Diarization. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1 (2012), 79–93.
33. K. Otsuka. 2011. Multimodal Conversation Scene Analysis for Understanding People's Communicative Behaviors in Face-to-Face Meetings. In *Human Interface and the Management of Information*. 171–179.
34. K. Otsuka, H. Sawada, and J. Yamato. 2007. Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "who responds to whom, when, and how?" from gaze, head gestures, and utterances. In *ICMI*. 255–262.
35. R. J. Rienks, R. Poppe, and D. Heylen. 2010. Differences in head orientation behavior for speakers and listeners: an experiment in a virtual environment. *ACM Transactions on Applied Perception* 7, 1 (2010), 2:1–2:13.
36. K. Ruhland, C. E Peters, S. Andrist, J. B Badler, N. I Badler, M. Gleicher, B. Mutlu, and R. McDonnell. 2015. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum* 34, 6 (2015), 299–326.
37. R. Stiefelhagen. 2002. Tracking Focus of Attention in Meetings. In *ICMI*. 273–280.
38. S. E. Tranter and D. A. Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 5 (2006), 1557–1565.
39. R. Vertegaal, G. van der Veer, and H. Vons. 2000. Effects of Gaze on Multiparty Mediated Communication. In *Graphics Interface*. Morgan Kaufmann, 95–102.